**LECTURE 1**

**INTRODUCTION TO STATA, DATA MANAGEMENT AND DESCRIPTIVE STATISTICS**

STATA is one of the most popular statistical software for analyzing socioeconomic databases (although any type of data can be analyzed using STATA). It is widely used by national and international statistical agencies, central banks, think tanks and academic researchers. For example, the Visitinps Scholar program, which gives researchers access to INPS administrative databases upon request, uses STATA software to enable data analysis.

How can you install STATA on your computer? One must request the software by opening a ticket on UNITO's Service Desk system. More information can be found at this link:

<https://www.unito.it/servizi/servizi-line/licenze-software-campus-di-ateneo>

There are excellent online resources that can be useful for learning how to use STATA and for finding solutions to particular problems. When you encounter a problem in using the software, or if you have doubts about how a particular operation can be performed, it is always a good idea to first try “asking Google.” In most cases, you will be directed to the solution of your problems! Some references to well-structured “online user manuals” are below:

<https://stats.idre.ucla.edu/stata/>

<https://www.stata.com/links/resources-for-learning-stata/>

In this lesson we begin to use STATA through a practical example. We will study some characteristics of labor and pension earnings in the Italian labor market. For this part of the lesson we use data provided by the Bank of Italy. In particular, we will study one of the archives of the “Indagine sui bilanci delle famiglie italiane”. The database that we will use during the lecture can be found in the file:

“rper.dta”

Please note, the file extension “.dta” always refers to dataset, it is the default format with which STATA opens and saves data.

Some references about the dataset we will analyze in this lesson:

* The excel file documentazione\_storico\_versione1.1 (provided in the course material) contains a description of the variables for each archive of the survey, including the rper.dta archive that we will use in this lecture
* The full database can be downloaded at the following link (this is potentially an interesting source for a master thesis…)

https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/distribuzione-microdati/index.html

1. **PRELIMINARY OPERATIONS**

1) Let's create a working folder in the computer memory related to this lesson.

- Let's create a folder in the computer memory (e.g. “C:\Desktop\lecture\_1”). This will be the working folder where we will save all the files related to this lesson.

- We copy the file “vote.dta” containing the database we will analyze into this working folder

2) Let's open the STATA software and start writing a do file

Immagine che contiene testo, schermata, software, Icona del computer

Descrizione generata automaticamente

The main window that appears in STATA contains the results of the commands we execute with this software (results window).

Commands in STATA can be executed by writing them to the command window (by default this is the window located at the bottom, below the results window). A better option is to write all commands in a “.do” file (this is a file containing a list of commands, which can be saved and edited after each session like a normal text file). Some information about “.do” files (which we will also call “do files”):

- To open a new do file, press the “do-file editor” button located on the toolbar at the top of STATA

Immagine che contiene testo

Descrizione generata automaticamente

- In a do file, each command must be written in a separate line. In addition, you must respect upper/lower case letters and the presence of blanks between the words that make up a command line

- To execute a command (or commands) written in a do-file, select the line(s) with the command(s) you want to execute, then press the “do” button located on the toolbar at the top of the do-file. If you want to execute the entire list of commands written in a do-file, simply press the “do” button without selecting any particular line.

- You can write a comment in a do-file by starting a line with a “\*” sign (comments are annotations you might want to include to remind you what you are doing in the do-file; STATA understands that these are not commands). If you want to write comments using more than one line, you must instead enclose the comment between a “/\*” sign and a “\*/” sign. Comments in a do file are highlighted in green by default.

- All the commands we will discuss later in this lesson can be written to and executed on the do file. At the end of the lesson remember to save the do file using the save button (just as you would save any file). This way you will not lose the work you have done and you will not have to rewrite all the commands each time.

- REMEMBER: The order in which I run the individual commands is important. Analyzing a database with STATA is a sequential operation: you generally start by opening the database (loading it into STATA's memory), continue by creating/editing database variables, and finally perform certain statistical analyses. You cannot create/edit a variable if we have not first opened the database, just as you cannot do a statistical analysis of a variable if that variable is not already in the database/not created.

3) We set the lesson folder as the default destination where STATA search for saves files

We write in the do file and execute the command **cd “C:\Desktop\lection\_1”** (the directory inside the quotes depends on where you saved the lesson folder on your computer). This command tells STATA that all the files we want to use and all the files we want to save are in this folder.

4) Open a .log file

We run the **cap log close** command to close any log files already open. We then run the command **log using lecture1.log, replace** This command creates a file called lecture1.log in our working directory. This file will be a text file that reproduces all the results of our commands (it records everything that appears in the STATA results window). It is very useful to save the results of your analyses, this way we do not have to run the do file every time we want to review these results, just open the log file we saved!

At the end of the analyses in this lesson, we will enter the **log close** command in the do file to tell STATA that the analysis is over and we no longer want to continue recording the results in the log file.

5) Some hints on the command structure and the help command

So far we have executed only three commands:

**cd “C:\Desktop/lecture\_1”**

**cap log close**

**log using lecture1.log, replace**

Typically, the first word (or words) of a command indicates the operation we want to perform. For example, “**cd**” stands for “change directory,” meaning “change the directory where STATA looks for and saves files that are used.” The second part of the command **“C:\Desktop\lecture\_1”** tells STATA what directory we want to be used, and this part of the command depends on the specific needs of my analysis (e.g., a future lecture will use a different working directory and so this part of the command will be different). **log using** tells STATA that we want to create a log file that records results, the second part (**lecture1.log**) indicates the name and format with which we want the log file to be saved. This second part of the command could have a different name depending on my preference. The last part of the command (**, replace**) is preceded by a comma. Generally, the command parts preceded by a comma are optional, that is, they specify certain ways in which we want the command to be executed. In the case of **, replace** we are indicating to STATA that if a file named lecture1.log already existed in the folder “C:\Desktop\_1,” we want to over-write this file. If, on the other hand, I wanted to add items below an existing log file with the same name, I would have to use the option **, append**

Note also that the order in which I run the two commands is important. By running **cd “C:\Desktop\lection\_1”** before **log using lecture1.log, replace** STATA will save the lecture1.log file in the C:\Desktop\lection\_1 folder. If I ran the commands in reverse order, STATA would save the lecture1.log file in the default destination that is set every time I open STATA!

How many commands exist in STATA? Several thousands; it is impossible to know them all. Each command also often has several options that allow you to choose from different settings. We generally learn the meaning of each command by trying to use it at the time we need it, and in these lessons we will follow that approach.

In case of doubt about the syntax and meaning of a command, STATA also contains a help feature, called help. Let's try running the command **help log** This command opens a new window that contains a description of the “log” command and all its features. The help command can be used in relation to any STATA command, for example try running the command **help cd** Now let's delete the help log and help cd commands from the do file. These are not commands that produce results or otherwise necessary for our data analysis work, and so we do not want to keep them in the do file where we will save the commands needed to replicate the statistical analysis in this lesson.

**2. EXPLORATORY ANALYSIS OF THE RPER.DTA DATABASE**

You may want to accompany these instructions by reading the lesson do file, which contains each of the commands discussed here. You should try running each of these commands on STATA as they are discussed and observe the results produced by each command (which appear in the results window and are recorded in the lesson log file). Remember that the order in which you run the commands is important (and running the same command twice can in some cases result in an error message, highlighted in red in the results window).

a) We load (open) the reper.dta file into STATA memory by running the “**use rper.dta, clear**” command.

b) The “**describe**” (“**des**”) command provides a description of the dataset (number of observations, number of variables, type and label of variables included in the dataset). The “**summarize**” (“**sum**”) command provides some descriptive statistics of the variables (number of observations net of any empty cells, mean, standard deviation, minimum and maximum value).

Note: summarize produces useful results only if the variable whose descriptives we calculate is numeric. If a variable contains letters or otherwise non-numeric characters, STATA will not compute its mean or other statistics that require numeric values to be computed (this is also true for other commands besides summarize). In the language of STATA, non-numeric variables are called strings.

(c) We can view our data by running “**browse**.” Alternatively, we can edit them by running “**edit**” Both commands involve opening a new window with an excel-like data table containing the database we are analyzing. We can close this window and continue the analysis.

3. **SELECTION OF VARIABLES AND OBSERVATIONS OF INTEREST, RENAMING AND LABELLING OF VARIABLES**

Please check the .log and .do files of the lectures for an explanation of each command

4. **HYPOTHESIS TESTING**

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

5. **HYPOTHESIS TESTING FOR CATEGORICAL VARIABLES**

Please check the .log and .do files of the lectures for an explanation of each command

6. **LONGITUDINAL ANALYSIS: INDIVIDUAL INCOME GROWTH**

Please check the .log and .do files of the lectures for an explanation of each command

7. **LONGITUDINAL ANALYSIS: THE REPLACEMENT RATE**

**Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente**